Analysis of Economics Data
Chapter 7: Statistical Inference for Bivariate Regression

© A. Colin Cameron
Univ. of Calif. Davis

November 2022

CHAPTER 7: Statistical Inference for Bivariate Regression

- Recall univariate
  - sample mean $\bar{x}$ estimates population mean μ
  - under suitable assumptions $t = \frac{\bar{x}-\mu}{se(\bar{x})}$ is a draw from $T(n-1)$
  - use this as basis for confidence intervals and hypothesis tests on μ.
- Now for bivariate regression
  - sample slope coefficient b_2 estimates population slope coefficient β_2
  - under suitable assumptions $t = \frac{b_2-\beta_2}{se(b_2)}$ is a draw from $T(n-2)$
  - use this as basis for confidence interval and hypothesis tests on β_2.

Outline

1. Example: House Price and Size
2. The t Statistic
3. Confidence Intervals
4. Tests of Statistical Significance
5. Two-Sided Hypothesis Tests
6. One-Sided Hypothesis Tests
7. Robust Standard Errors
8. Examples

Dataset: HOUSE.

7.1 Example: House Price and Size

- Key regression output for statistical inference with $n = 29$:

Variable	Coefficient	Standard Error	t statistic	p value	95% conf. interval	
Size	73.77	11.17	6.60	0.000	50.84	96.70
Intercept	115017.30	21489.36	5.35	0.000	70924.76	159109.8

- $\widehat{price} = b_1 + b_2 size$ is an estimate of $price = \beta_1 + \beta_2 size$.
- Coefficient of `Size`
  - $b_2 = 73.77$ is least squares estimate of slope β_2
- Standard error of `Size`
  - the estimated standard deviation of b_2
  - the **default standard error** of b_2 equals 11.17.
  - (later: alternative **heteroskedastic-robust standard errors**).

Example (continued)

- We have with $n = 29$:

Variable	Coefficient	Standard Error	t statistic	p value	95% conf. interval	
Size	73.77	11.17	6.60	0.000	50.84	96.70
Intercept	115017.30	21489.36	5.35	0.000	70924.76	159109.8

- Confidence interval for size
  - 95% confidence interval for β_2
  - is $b_2 \pm t_{27,.025} \times se(b_2) = (50.84, 96.70)$.
- t statistic of `Size` tests whether there is any relationship
  - is for test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$
  - in general $t =$ (estimate− hypothesized value)/standard error.
  - $t_2 = b_2 / se(b_2) = 73.77/11.17 = 6.60$.
- p value of `Size`
  - is p-value for a two sided test
  - $p_2 = \Pr[|T_{27}| > |6.60|] = 0.00$.

7.2 The t Statistic

- The statistical inference problem
  - **Sample**: $\widehat{y} = b_1 + b_2 x$ where b_1 and b_2 are least squares estimates
  - **Population**: $\mathrm{E}[y|x] = \beta_1 + \beta_2 x$ and $y = \beta_1 + \beta_2 x + u$.
  - **Estimators**: b_1 and b_2 are estimators of β_1 and β_2.
- Goal
  - **inference on the slope parameter** β_2.
- This is based on a $T(n-2)$ distributed statistic

$$T = \frac{\text{estimate} - \text{parameter}}{\text{standard error}} = \frac{b_2 - \beta_2}{se(b_2)} \sim T(n-2).$$

Why use the T(n-2) Distribution?

- Make assumptions 1-4 given in the next slide.
  - then $\text{Var}[b_2] = \sigma_u^2 / \sum_{i=1}^n (x_i - \bar{x})^2$.
- But we don't know σ_u^2
  - we replace it with the estimate $s_e^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \widehat{y}_i)^2$.
- This leads to noise in $\{se(b_2)\}^2 = s_e^2 / \sum_{i=1}^n (x_i - \bar{x})^2$
  - so the statistic $T = (b_2 - \beta_2)/se(b_2)$ is better approximated by $T(n-2)$ than by $N(0,1)$.
- The $T(n-2)$ distribution
  - is the exact distribution if additionally the errors u_i are normally distributed
  - otherwise it is an approximation, one that computer packages use.

Model Assumptions

- **Data assumption** is that there is variation in the sample regressors so that $\sum_{i=1}^{n}(x_i - \bar{x})^2 = 0$.
- **Population assumptions 1-4**
  - **1.** The **population model** is $y = \beta_1 + \beta_2 x + u$.
  - **2.** The **error has mean zero conditional on x**: $\mathrm{E}[u_i | x_i] = 0$.
  - **3.** The **error has constant variance conditional on x**: $\mathrm{Var}[u_i | x_i] = \sigma_u^2$.
  - **4.** The **errors for different observations are statistically independent**: u_i is independent of u_j.
- Assumptions 1-2 imply a linear conditional mean and yield unbiased estimators

$$\mathrm{E}[y|x] = \beta_1 + \beta_2 x.$$

- Additional assumptions 3-4 yield the variance of estimators.

7.3 Confidence Interval for the Slope Parameter

- Recall: A 95 percent confidence interval is approximately

$$\text{estimate} \pm 2 \times \text{standard error}$$

  - here a 95% confidence interval is $b_2 \pm t_{n-2;.025} \times se(b_2)$.

- A $100(1-\alpha)$ **percent confidence interval for** β_2 is

$$b_2 \pm t_{n-2,\alpha/2} \times se(b_2),$$

  where

  - b_2 is the slope estimate
  - $se(b_2)$ is the standard error of b_2
  - $t_{n-2;\alpha/2}$ is the critical value in Stata using `invttail(n-2,`α`/2)`.

What Level of Confidence?

- There is no best choice of confidence level
    - most common choice is 95% (or 90% or 99%)
- Interpretation
    - the calculated 95% confidence interval for β_2 will correctly include β_2 95% of the time
    - if we had many samples and in each sample formed a 95% confidence interval, then 95% of these confidence intervals will include the true unknown β_2.

Example: House Price and Size

- For regress house price on house size a 95% confidence interval is

$$\begin{aligned} & b_2 \pm t_{n-2,\alpha/2} \times se(b_2) \\ = \; & 73.77 \pm t_{27,.025} \times 11.17 \\ = \; & 73.77 \pm 2.052 \times 11.17 \\ = \; & 73.77 \pm 22.93 \\ = \; & (50.84, 96.70). \end{aligned}$$

- This is directly given in computer output from regression.

7.4 Tests of Statistical Significance

- A regressor x has **no relationship** with y if $\beta_2 = 0$.
- A test of "**statistical significance**" is a two-sided test of whether $\beta_2 = 0$. So test

$$H_0 : \beta_2 = 0 \quad \text{against} \quad H_a : \beta_2 \neq 0.$$

- Test statistic is then

$$t = \frac{b_2}{se(b_2)} \sim T(n-2).$$

- Reject if $|t|$ is large as then $|b_2|$ is large
  - How large?
  - Large enough that the value of $|t|$ is a low probability event.
- Use either p value approach or critical value approach
  - reject at level 0.05 if $p = \Pr\{|T_{n-2}| > |t|\} < 0.05$
  - or equivalently reject at level 0.05 if $|t| > c = t_{n-2;.025}$.
- This method generalizes to other formulas for $se(b_2)$.

Example: House Price and Size

- For regress house price on house size with $n = 29$

$$t = \frac{b_2}{se(b_2)} = \frac{73.77}{11.17} = 6.60$$

- $p = \Pr[|T_{n-2}| > |t|] = \Pr[|T_{27}| > 6.60] = 0.000$
  - so reject $H_0 : \beta_2 = 0$ at significance level 0.05 as $p < 0.05$.
- $c = t_{n-2;.025} = t_{27,.025} = 2.052$
  - so reject H_0 at significance level 0.05 as $|t| = 6.60 > c$.
- Conclude that house size is statistically significant at level 0.05.

Economic Significance versus Statistical Significance

- A regressor is of **economic significance** if its coefficient is of large enough value for it to matter in practice
  - economic significance depends directly on b_2 and the context
- By contrast, **statistical significance** depends directly on t which is the ratio $b_2/se(b_2)$.
- With large samples $se(b_2) \longrightarrow 0$ as $n \longrightarrow \infty$
  - so we may find statistical significance
  - even if b_2 is so small that it is of little economic significance.

Tests based on the Correlation Coefficient

- An alternative way to measure statistical significance, used in many social sciences, uses the **correlation coefficient** $|r_{xy}|$.
- Then reject the null hypothesis of no association if $|r_{xy}|$ is sufficiently large
  - this gives similar results to tests based on $t = b_2/se(b_2)$ if default standard errors are used.
- Weaknesses of tests using the correlation coefficient
  - this method cannot relax assumptions 3-4
  - this method cannot be used if we wish to add additional regressors
  - and it tells little about economic significance.

7.5 Two-sided Hypothesis Tests

- A **two-sided test** on the slope coefficient is a test of

$$H_0 : \beta_2 = \beta_2^* \quad \text{against} \quad H_a : \beta_2 \neq \beta_2^*.$$

- Use t-statistic where $\beta_2 = \beta_2^*$. So compute

$$t = \frac{b_2 - \beta_2^*}{se(b_2)} \sim T(n-2).$$

- Reject if $|t|$ is large as then $|b_2 - \beta_2^*|$ is large
  - How large?
    - Large enough that such a large $|t|$ is a low probability event.
  - Use either p value approach or critical value approach.

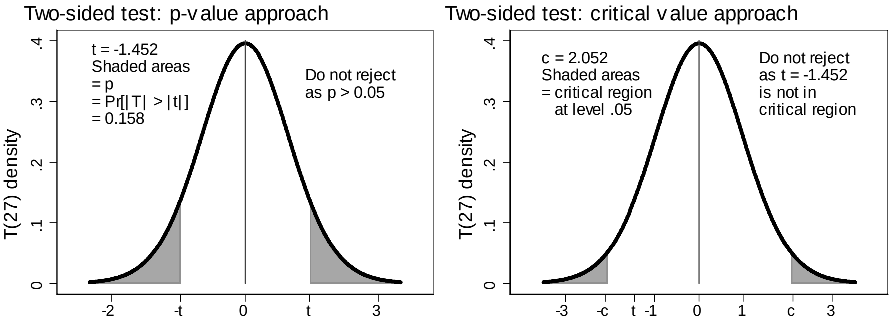
Example: House Price and Size

- For house price example with $\beta_2^* = 90$

$$t = \frac{b_2 - 90}{se(b_2)} = \frac{73.77 - 90}{11.17} = -1.452.$$

- p-value approach
  - $p = \Pr[|T_{27}| > |-1.452| = 0.158.$
  - do not reject H_0 at level 0.05 as $p = 0.158 > 0.05.$
- Critical value approach at level 0.05:
  - $c = t_{27;.025} = 2.052.$
  - do not reject H_0 at level 0.05 as $|t| = 1.452 < c = 2.052.$
- In either case we do not reject $H_0 : \beta_2 = 90$ against $H_a : \beta_2 \neq 90$ at level 0.05.
  - conclude that house price does not increase by $90 per square foot.

- p-value approach: Compute $p = \Pr[|T_{n-2}| > |t|]$.
- critical value approach: compute c so that reject if $|t| > c$.



Rejection using p-values

- **p-value approach** (at level $\alpha = 0.05$)
  - Assume that $\beta_2 = \beta_2^*$, i.e. H_0 is true.
  - Obtain the p-value
    - the probability (or significance level) of observing a $|T_{n-2}| \geq |t|$, where this probability is calculated under the assumption that $\beta_2 = \beta_2^*$.
  - If $p < 0.05$ then reject H_0
    - reason there was less than .05 chance of observing our t, given $\beta_2 = \beta_2^*$.

Rejection using Critical values

- **Critical value approach** (at level $\alpha = 0.05$)
  - Assume that $\beta_2 = \beta_2^*$, i.e. H_0 is true.
  - Find the critical value
    - the value c such that $\Pr[|T_{n-2}| \geq c] = 0.05$
  - If $|t| > c$ then reject H_0
    - reason: there was less than .05 chance of observing our t, given $\beta_2 = \beta_2^*$.

Relationship of Tests to Confidence Interval

- For a two-sided test of $H_0 : \beta_2 = \beta_2^*$
  - if the null hypothesis value β_2^* falls inside the $100(1-\alpha)$ percent confidence interval then do not reject H_0 at significance level α.
  - otherwise reject H_0 at significance level α.
- House example
  - 95% confidence interval for β_2 is (50.84, 96.70)
  - reject $H_0 : \beta_2 = 0$ at level 0.05 as the 95% confidence interval does not include 0.

7.6 One-sided Directional Hypothesis Tests

- **One-sided test** on the slope coefficient is a test of

Upper one-tailed alternative	$H_0 : \beta_2 \leq \beta_2^*$	against	$H_a : \beta_2 > \beta_2^*$
Lower one-tailed alternative	$H_0 : \beta_2 \geq \beta_2^*$	against	$H_a : \beta_2 < \beta_2^*$

- The statement being tested is specified to be the alternative hypothesis.
- Use same t-statistic as in two-sided case. So

$$t = \frac{b_2 - \beta_2^*}{se(b_2)} \sim T(n-2).$$

- What will differ is the rejection region
  - For $H_0 : \beta_2 \leq \beta_2^*$ against $H_a : \beta_2 > \beta_2^*$ reject in the right tail
    - ★ $p = \Pr[T_{n-2} > t]$
  - For $H_0 : \beta_2 \geq \beta_2^*$ against $H_a : \beta_2 < \beta_2^*$ reject in the left tail
    - ★ $p = \Pr[T_{n-2} < t]$.

Example: House Price and Size

- House price example suppose claim is that house price rises by less than \$90 per square foot, i.e. $\beta_2 < 90$.

- Test $H_0 : \beta_2 \geq 90$ against $H_a : \beta_2 < 90$ (lower tailed alternative).

$$t = \frac{b_2 - 90}{se(b_2)} = \frac{73.77 - 90}{11.17} = -1.452.$$

- p-value approach:
  - $p = \Pr[T_{27} < t] = \Pr[T_{27} < -1.452]$
    $= \Pr[T_{27} > 1.452] = \texttt{ttail}(27, 1.452) = 0.079 < 0.05.$
    - ★ where we have used the symmetry of the t distribution.
- Critical value approach at level 0.05:
  - $c = -t_{27,.05} = -\texttt{invttail}(27, .05) = -1.70$ and $t \nless -1.70$.
- In either case we do not reject $H_0 : \beta_2 \geq 90$ at significance level 0.05.
- At level 0.05 there is not enough evidence to support the claim
  - note that the claim would be supported if we tested at level 0.10.

Computer generated t-statistic

- Computer gives a t-statistic
  - this is $t = b_2 / se(b_2)$
  - suitable for testing $\beta_2 = 0$.
- Computer gives a p-value
  - this is for a two-sided test of $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$.
- For a one-sided test of statistical significance
  - if b_2 is of the expected sign then halve the printed p-value.
  - if b_2 is not of the expected sign then reject since $p > 0.5$
- Example: if expect $\beta_2 > 0$ then upper tailed alternative test
  - test $H_0 : \beta_2 \leq 0$ against $H_a : \beta_2 > 0$ at level .05
  - if $b_2 > 0$ then halve the printed p value and reject H_0 if this is less than .05
  - if $b_2 < 0$ we will not reject H_0 i.e. conclude β_2 is not greater than zero.

7.7 Robust Standard Errors

- **Default standard errors** (and associated t statistics, p values and confidence intervals) make assumptions 1-4
  - called **default** because this is what computer automatically computes
- **Robust standard errors**
  - Keep assumptions 1-2
  - Relax assumptions 3-4 in three common ways depending on data type
  - Are commonly-used in practice.
- In each case get an alternative formula for $se(b_2)$, say $se_{rob}(b_2)$
- Then base inference on

$$t = \frac{b_2 - \beta_2}{se_{rob}(b_2)}.$$

Heteroskedastic Robust Standard Errors

- Relax assumption 3 that all errors have the same variance
  - called the assumption of **homoskedastic errors**.
- Instead allow $\text{Var}[u_i|\mathbf{x}_i] = \sigma_i^2$ which varies with i
  - called **heteroskedastic errors**.
- This is the standard assumption in modern econometrics.
- Then the heteroskedasticity-robust standard error for b_2 is

$$se_{het}(b_2) = \frac{\sqrt{\sum_{i=1}^{n} e_i^2 (x_i - \bar{x})^2}}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \neq \frac{s_e}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}}.$$

- Then $t = (b_2 - \beta_2)/se_{het}(b_2)$ is viewed as $T(n-2)$ distributed.

Example: House Price and Size

- For the house price and size example
  - default standard errors
    - 11.17 and 21,489 for the slope and intercept
  - heteroskedastic-robust standard errors
    - 11.33 and 20,928 for the slope and intercept
- Confidence interval using heteroskedastic-robust standard errors
  - $73.77 \pm t_{27,.025} \times 11.333 = (50.33, 97.02)$ compared t0 $(50.84, 96.70)$
- Test $H_0 : \beta_2 = 0$ against $H_a : \beta_2 \neq 0$

$$t = \frac{b_2}{se(b_2)} = \frac{73.77 - 0}{11.33} = 6.51 \text{ compared to } 6.60.$$

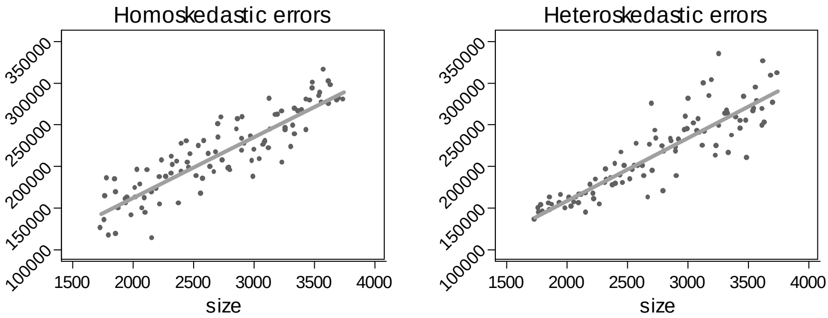
Simulation Example of Heteroskedastic Errors

- Generate 100 observations as follows
  - size varies from 1700 to 3700 plus some random noise
  - price = 11500 + 74*size + zero-mean error
  - (1) error is homoskedastic $u_i \sim N(0, 23500^2)$
  - (2) error is heteroskedastic $u_i \sim \frac{(\text{size}_i - 1700)}{1400} \times N(0, 23500^2)$
    - ★ this error has variance $\left\{\frac{(\text{size}_i - 1700)}{1400}\right\}^2 \times 23500^2$ that differs across i
- Stata code

```
set obs 100
generate size = 1700 + 20*_n + runiform(0,50)
generate uhomosked = rnormal(0,23500)
generate price = 11500 + 74*size + uhomosked
scatter price size || lfit price size
generate uheterosked = ((size-1500)/1400)*rnormal(0,23500)
generate price2 = 11500 + 74*size + uheterosked
scatter price2 size || lfit price size
```

Simulation Example (continued)

- First panel: homoskedastic errors are evenly distributed around the regression line.
- Second panel: heteroskedastic errors scattering around the regression line varies with the level of the regressor
  - in this case increasing with regressor `size`.



Other Robust Standard Errors

- For time series data where model **errors may be correlated over time**
  - use HAC robust.
- For data in clusters (or groups) where **errors are correlated within cluster** but are uncorrelated across clusters
  - people in villages, students in schools, individuals in families, ...
  - panel data on many individuals over time
  - use cluster robust.
- These robust standard errors are presented in chapter 12.1.
- An essential part of any regression analysis is knowing which particular robust standard error method should be used.

Key Stata Commands

```
clear
use AED_HOUSE.DTA
regress price size
regress price size, level(99)
* Following gives F = t-squared and correct p-value
test size = 90
regress price size, vce(robust)
```

Some in-class Exercises

1. We obtain fitted model $\widehat{y} = \underset{(1.5)}{3.0} + \underset{(2.0)}{5.0} \times x$, $R^2 = 0.32$, $s_e = 4.0$, $n = 200$. Provide an approximate 95% confidence interval for the population slope parameter.
2. Test the claim that the population slope equals 2 at the 5% significance level.
3. Which of assumptions 1-4 need changing if model errors are heteroskedastic?